Extracting Actionable Insights from User Movie Reviews

A report submitted in partial fulfillment of the requirements for Colloquium (BCCS-4105)

Bachelor of Technology
in
Computer Science and Engineering

by

Bhavana Mekala 2021BCS-41

विश्वजीवनामृतं ज्ञानम्

ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT
GWALIOR - 474015

July 2024

CANDIDATES DECLARATION

We hereby certify that the work, which is being presented in the report, entitled **Actionable Insights**, in partial fulfillment of the requirement for Colloquium (BCCS-4105) for **Bachelor of Technology in Computer Science and Engineering** and submitted to the institution is an authentic record of our own work carried out during the period *May 2024* to *July 2024* under the supervision of **Dr. Santosh Singh Rathore**. We also cited the reference about the text(s) from where they have been taken.

Date: Signature of the Candidate

This is to certify that the above statement made by the candidates is correct to the best of my knowledge.

Date: Signature of the Supervisor(s)

ABSTRACT

This project presents an in-depth analysis of user reviews from the IMDB dataset to extract actionable insights using advanced natural language processing (NLP) and machine learning techniques. The primary focus is on performing sentiment analysis to classify reviews as positive or negative and then applying topic modeling on negative reviews to uncover underlying themes. The sentiment analysis was conducted using traditional machine learning models, including Complement Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes, as well as a deep learning approach using Bidirectional Long Short-Term Memory (BiLSTM) networks. For topic modeling, Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) were utilized. The deep learning model achieved an 85.89% accuracy in sentiment classification, and the LSA model produced a coherence score of 0.79 for topics extracted from negative reviews. These findings highlight the potential of combining traditional and deep learning methods to effectively analyze large volumes of textual data, providing valuable insights for stakeholders.

Keywords: Semantic Analysis, Topic Modelling.

ACKNOWLEDGEMENTS

We are highly indebted to **Dr. Santosh Singh Rathore**, and are obliged for giving us the autonomy of functioning and experimenting with ideas. We would like to take this opportunity to express our profound gratitude to him not only for his academic guidance but also for their personal interest in our project and constant support coupled with confidence boosting and motivating sessions which proved very fruitful and were instrumental in infusing self-assurance and trust within us. The nurturing and blossoming of the present work is mainly due to their valuable guidance, suggestions, astute judgment, constructive criticism and an eye for perfection. Our mentor always answered myriad of our doubts with smiling graciousness and prodigious patience, never letting us feel that we are novices by always lending an ear to our views, appreciating and improving them and by giving us a free hand in our project. It's only because of his overwhelming interest and helpful attitude, the present work has attained the stage it has.

Finally, we are grateful to our Institution and colleagues whose constant encouragement served to renew our spirit, refocus our attention and energy and helped us in carrying out this work.

Bhavana Mekala

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

User feedback, particularly in the form of online reviews, has become an indispensable resource in shaping perceptions and influencing decisions across various industries. For platforms like IMDB, which hosts an extensive repository of user-generated content, effectively analyzing and extracting insights from these reviews is crucial. Not only can this process enhance user experience and content offerings, but it can also guide strategic decisions and drive platform improvements. However, the unstructured nature of textual data presents a significant challenge. Deriving actionable insights from such data requires the application of robust Natural Language Processing (NLP) and machine learning techniques.

Motivated by the potential to improve the IMDB platform and better understand user sentiments, this project was undertaken to tackle the complex task of processing and analyzing large-scale user reviews from the IMDB dataset. The primary objective was to classify the sentiment of these reviews and further explore the specific concerns and themes expressed within negative reviews. By addressing these challenges, the project aimed to provide valuable insights that could inform decisions related to content development, user engagement, and overall platform enhancement.

The project adopted a comprehensive, multi-step approach that began with text preprocessing, a critical step in preparing the data for analysis. This phase involved cleaning the text by removing noise such as special characters, HTML tags, and URLs, normalizing the text by converting it to lowercase, and tokenizing the words. Additional preprocessing steps included the removal of stopwords, lemmatization to reduce words to their base forms, and filtering out common and infrequent words to focus on meaningful content. These preprocessing steps are well-established in the field of sentiment analysis and are known to significantly improve the performance of subsequent models.

Following preprocessing, feature extraction was conducted using two popular techniques: CountVectorizer and Term Frequency-Inverse Document Frequency (TF-IDF). CountVectorizer converts the text into a matrix of token counts, providing a simple yet effective way to quantify the words in the dataset. TF-IDF, on the other hand, adjusts the frequency of words by considering how commonly they appear across all documents, thus helping to identify words that are particularly important in specific reviews.

For the sentiment classification task, the project explored both traditional machine learning methods and advanced deep learning techniques. Three variants of Naive Bayes models—MultinomialNB, BernoulliNB, and ComplementNB—were tested, each offering unique advantages depending on the distribution and characteristics of the text data. However, recognizing the limitations of traditional models in capturing complex patterns in text, a Bidirectional Long Short-Term Memory (BiLSTM) model was also employed. The BiLSTM model was chosen for its ability to understand the context of words within a sequence, an essential feature for accurately determining sentiment in text. By considering both past and future word sequences, the BiLSTM model is particularly effective in handling the nuances of language, making it a powerful tool for sentiment analysis.

Once the reviews were classified into positive and negative sentiments, the project delved deeper into the negative reviews to extract further insights. This was achieved through topic modeling, a technique designed to uncover hidden themes and patterns within large collections of text. Two methods were employed for this purpose: Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA). LDA is a generative statistical model that assumes each document is a mixture of several topics, with each topic being a distribution over words. It is widely used in NLP for discovering abstract topics in a corpus. LSA, on the other hand, uses singular value decomposition to identify patterns in the relationships between terms and concepts, providing a more mathematical approach to topic modeling.The coherence score was used to measure the effectiveness of these topic models. This score evaluates how well the identified topics represent coherent and meaningful themes.

Previous research in sentiment analysis and topic modeling has demonstrated the importance of meticulous preprocessing and the selection of appropriate models. Studies have consistently shown that preprocessing steps such as noise removal, stopword elimination, and lemmatization are critical for enhancing the accuracy and reliability of sentiment classification models. Moreover, topic modeling techniques like LDA and LSA have proven effective in revealing underlying themes in text data, offering valuable insights that can drive decision-making.

1.2 Literature Review

While the problem of topic extraction has been addressed in the past, in recent years research is relating topic extraction with sentiment analysis. In their work [1] the authors are interested to mine users' opinions on Weblogs, analyzing the sentiments for subtopics. In their approach the authors propose a probabilistic mixture model called Topic Sentiment Mixture (TSM) where words are sampled by a mixture model of background language, topic language and two sentiment language models. They present a mechanism for extracting subtopics, associating with every subtopic a positive or a negative sentiment and how the opinions over a topic change over time. Their approach does not use LDA and the sentiment model is applied as a post-processing step to the topic discovery.

In their work [2] the authors present JST, a method that is using a weakly- supervised approach to draw words taking into consideration both topics and sentiment labels from a corpus of documents thus extending LDA. As a result, JST performs document level sentiment classification where topics and senti- ments are detected simultaneously while it can extract sentiment oriented topics effectively evaluating the sentiment of each topic. In [3] the authors extend JST proposing the Sentiment LDA, where sentiment labels are associated to topics instead of documents and introduce sentiment dependency in their calculations.

In more recent work,[4], the authors argue that the sentiment should not be used to influence the topic as done in JST but sentiment polarities as well as topics of text should be analyzed at the same time. They propose Double Latent Dirichlet Allocation (DLDA) for sentiment analysis in short texts. A review of on LDA-based topic extraction in sentiment analysis is presented in [5]. As already presented, there is a diversity of approaches in the literature regarding the extraction of topics and their associated sentiments.

1.3 Objective

The primary objective of this project is to develop an automated system capable of extracting actionable insights from large volumes of user feedback, specifically focusing on the IMDB movie review dataset. The key objectives include:

1. To develop a sentiment classification model which classifies the movie reviews into positive and negative sentiments.

2. To apply topic modeling techniques to the subset of negative reviews in order to uncover recurring themes and concerns expressed by users.

Perform Sentiment Analysis to classify user reviews as positive or negative using both traditional machine learning algorithms (e.g., Complement Naive Bayes, Multinomial Naive Bayes, Bernoulli Naive Bayes) and a deep learning model (Bidirectional Long Short-Term Memory - BiLSTM) and then apply topic modeling techniques such as Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA) to the subset of negative reviews in order to uncover recurring themes and concerns expressed by users.

CHAPTER 2

Methodology

In this section, I detail the implementation of my project, where I conducted sentiment analysis to classify user reviews as positive or negative. I employed traditional machine learning methods: Complement Naive Bayes, Multinomial Naive Bayes, and Bernoulli Naive Bayes alongside a deep learning model, Bidirectional LSTM (BiLSTM). Subsequently, I applied topic modeling techniques, including Latent Dirichlet Allocation (LDA) and Latent Semantic Analysis (LSA), to the negative reviews to identify recurring themes and concerns.

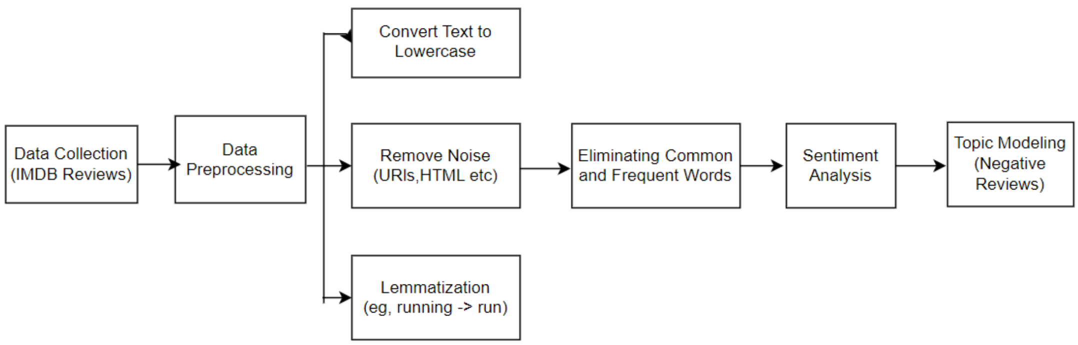


Figure 2.1: Flowchart of the project

2.0.1 Data preprocessing

Data preprocessing is a foundational step in text analysis, crucial for ensuring that the data is clean, uniform, and ready for analysis. The preprocessing phase in this project involves several key steps:

1. Converting Text to Lowercase: All text is converted to lowercase to ensure consistency across the dataset. This step is essential because it prevents the same word with different cases (e.g., "Film" and "film") from being treated as different entities, thereby reducing redundancy and improving analysis accuracy.

2. Removing Noise: To enhance the quality of the data, various forms of noise, such as URLs, digits, HTML tags, punctuation, and stopwords, are removed. URLs and HTML tags often do not contribute to the meaning of the text and can interfere with the analysis. Punctuation and digits are similarly irrelevant in sentiment analysis and topic modeling. Stopwords (common words like "and," "the," and "is") are removed because they are frequently used but carry little semantic weight in isolation.

3. Lemmatization: Words are reduced to their base or root form (lemmatization) to consolidate different forms of the same word. For instance, "running," "ran," and "runs" would all be reduced to "run." This step is crucial for ensuring that variations of a word are treated as a single entity, improving the consistency and quality of the dataset.

4. Eliminating Common and Frequent Words: Words that appear too frequently and do not add significant meaning to the text are removed. This includes domain-specific jargon or overly common terms that might dilute the focus of the analysis. By eliminating these words, the dataset becomes more refined, allowing for more meaningful insights to be drawn from the remaining content.

2.0.2 Sentiment Analysis

After the data has been preprocessed, sentiment analysis is performed to classify the reviews into positive and negative sentiments. This step is critical for understanding the overall user sentiment toward the films. The sentiment analysis model, trained on a large corpus of labeled data, analyzes the words and phrases used in the reviews to determine their sentiment.Reviews classified as positive reflect satisfaction, enjoyment, or approval of the films. These reviews often highlight aspects such as good acting, compelling storylines, or impressive visuals.Reviews classified as negative indicate dissatisfaction, disappointment, or disapproval. These reviews are particularly valuable for identifying issues or areas for improvement in the films.This binary classification of reviews into positive and negative categories helps in quantifying the overall sentiment and provides a basis for further analysis.

2.0.3 Topic Modeling

Following sentiment analysis, topic modeling is applied specifically to the negative reviews. The goal of this step is to identify common themes and topics within the negative feedback, providing deeper insights into the issues highlighted by users.Topic modeling helps in uncovering recurring themes in the negative reviews, such as complaints about the plot, acting, or special effects. By grouping related words and phrases into topics, this technique reveals the underlying concerns and areas where the films may be falling short.

The project employs two different approaches to sentiment analysis followed by topic modeling, allowing for a comparison of their effectiveness. The identified topics are evaluated using the coherence score, which measures the semantic similarity between words in a topic. A higher coherence score indicates that the topics are more meaningful and relevant to the context of the reviews.

2.1 Approach - 1

2.1.1 Sentiment Analysis by Naive Bayes

Naive Bayes is a probabilistic classifier based on Bayes' theorem, which assumes independence between predictors. Bayes Theorem is a mathematical formula used to determine the conditional probability of events. With the help of Bayes Theorem We can predict whether a movie review is positive or negative by choosing the class with the highest probability. Each feature in the text is considered independent of others.Features are words in the review. A word not seen in training data might make P(word/class) = 0. To handle this use Laplace smoothing to ensure no probability is zero.
P(word/class) = (count of word in class + 1) / (total words in class + V)
V is the unique word in the corpus.

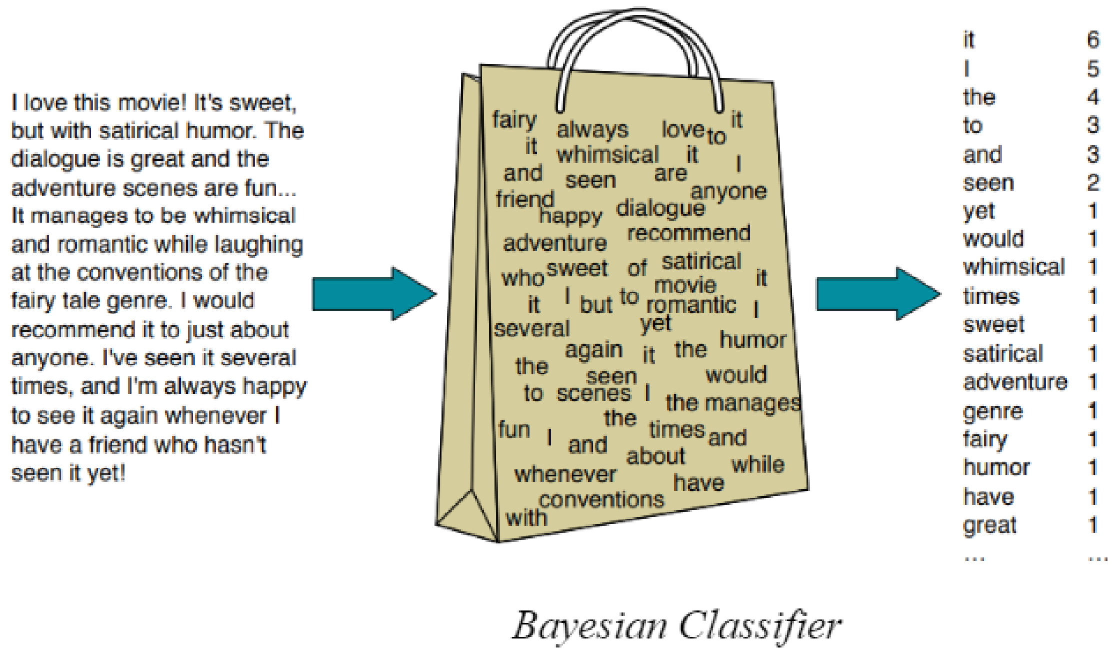
I have used three variants of Naive Bayes:

2.1.2 Multinomial Naive Bayes (MNB)

Multinomial Naive Bayes (Bayesian Classifier) is used for discrete data, particularly word counts in text. Assumes features are the frequencies or counts of words. In this project, this model is used to classify reviews based on the frequency of words. Widely applied for document classification and sentiment analysis.

2.1.3 Bernoulli Naive Bayes (BNB)

Bernoulli Naive Bayes is used for binary data, representing the presence or absence of words.Assumes features are binary indicators (0 or 1) of whether a word occurs in a document. In this project, this model looks at the presence or absence of specific words to classify reviews. Effective for document classification tasks that use binary features and spam detection.



[6] Daniel Jurafsky and James H. Martin. Speech and Language Processing. Draft of August 20, 2024, 2024. Copyright © 2024. All rights reserved.

2.1.4 Complement Naive Bayes (CNB)

Designed to address imbalances in the class distribution. A variant of Multinomial Naive Bayes, this model adjusts the calculation to complement the existing model, reducing bias towards majority classes. In this project, it helps improve performance on imbalanced datasets. Useful in text classification where class imbalances are a concern.

2.1.5 LDA for Topic Modeling

Latent Dirichlet Allocation (LDA) is a generative statistical model that identifies topics within a collection of documents. We import the Gensim library, which is commonly used for topic modeling and NLP tasks. Its corpora module helps in creating a dictionary and corpus needed for LDA. Next, we prepare the data for LDA by creating a dictionary from our preprocessed negative reviews. This dictionary maps each unique word to a unique integer ID, which is crucial for the model. For example, 'bad' might be mapped to 0, 'movie' to 1, and so on. Following this, we convert each document into a bag-of-words format using the 'doc2bow' method. It counts how many times each word appears in the document and outputs a list of tuples. Each tuple consists of the word ID and its count in the document, resulting in a corpus that captures the frequency of each word in the documents. For example, for the review "bad movie not worth time", the BoW representation might be [(0, 1), (1, 1), (2, 1), (3, 1), (4, 1)]. Once our data is prepared, we fit the LDA model using the 'LdaMulticore' function, which allows for parallel processing to speed up computations . We specify the number of iterations over the corpus using the 'passes' parameter and utilize 'workers' for parallel processing. Once the model is trained, we can extract the topics and the top words for each topic.

2.2 Approach - 2

2.2.1 Sentiment Analysis by Bidirectional LSTM

Bidirectional LSTM(Long Short -Term Memory) networks are effective for sequence prediction problems because they consider both past and future context. This allows for better understanding of sentiment in user reviews. We import libraries TensorFlow and Keras for building the neural network.

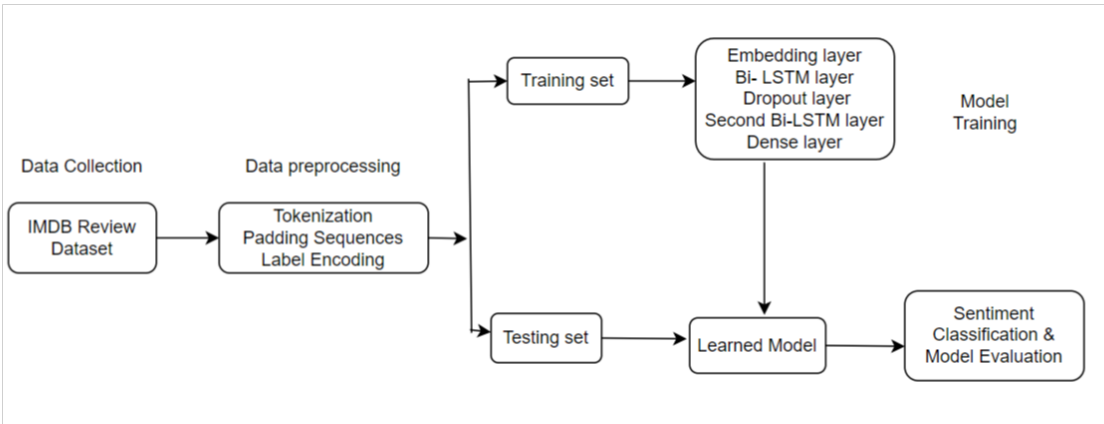


Figure 2.2: Flowchart of Sentiment Analysis by Bidirectional LSTM

We initialize a Tokenizer and fit it on the review texts. This transforms the text data into sequences of integers, where each integer represents a word in the vocabulary. We then pad the sequences to ensure uniform input lengths for the model. Using LabelEncoder, we transform the sentiment labels (e.g., positive or negative) into numerical format, which is necessary for training the model. We construct a sequential model with the following layers: 1. Embedding Layer: This layer maps the integer sequences to dense vectors of fixed size (embedding dimension). 2. Bidirectional LSTM Layer: This layer processes the input sequences in both forward and backward directions to capture context from both sides. 3. Dropout Layer: This layer randomly drops some neurons during training to prevent overfitting. 4. Second Bidirectional LSTM Layer: Another LSTM layer to further learn complex patterns in the data. 5. Dense Layer: This layer produces the final output with a sigmoid activation function, indicating whether the sentiment is positive or negative. After training, we use the model to make predictions on the test set. We convert the predicted probabilities into binary class labels (0 or 1) based on a threshold of 0.5.

2.2.2 LSA for Topic Modeling

Latent Semantic Analysis (LSA) is a technique that reduces dimensionality of text data and extracts topics from the dataset. We import TfidfVectorizer to convert a collection of text documents into a matrix of TF-IDF features. TF-IDF (Term Frequency-Inverse Document Frequency) is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus. It balances the frequency of words (Term Frequency) with how common they are across all documents (Inverse Document Frequency). 1. Term Frequency (TF): How often a word appears in a document. 2. Inverse Document Frequency (IDF): How common or rare a word is across all documents. Truncated SVD is used to reduce the dimensionality of the TF-IDF matrix and extract topics. Next, we initialize the TfidfVectorizer and use it to convert our collection of negative reviews into a TF-IDF matrix. Each document is represented as a row, and each term corresponds to a column, filled with TF-IDF scores. We retrieve the feature names (terms) from the TF-IDF vectorizer to access the actual words corresponding to the indices in the TF-IDF matrix.To understand the topics derived from the LSA model, we loop through each topic and find the top 10 terms. We do this by sorting the components of each topic and selecting the indices of the top terms.

CHAPTER 3

RESULTS

3.0.1 Results of Sentiment Analysis

The below table presents the outcomes of Sentiment Analysis conducted using both Naive Bayes methods and a Bidirectional LSTM model. Among the different Naive Bayes approaches, the BernoulliNB model demonstrates superior performance when compared to the MultinomialNB and ComplementNB models. However, it's worth noting that the Bidirectional LSTM model achieves even better results, surpassing BernoulliNB with an impressive accuracy of 85.89%. This highlights the effectiveness of the Bidirectional LSTM model in delivering more accurate sentiment analysis.

Table 3.1: Results of Sentiment Analysis

Sentiment Analysis	Accuracy	Precision	Recall	F1Ratio
ComplementNB	0.8182	0.8211	0.8211	0.8228
MultinomialNB	0.8202	0.8211	0.8245	0.8211
BernoulliNB	0.8285	0.8415	0.8247	0.8330
BiDirectional LSTM	0.8589	0.8496	0.8751	0.8621

3.0.2 Topics generated by LDA model

```
(0, '0.004*"jessica" + 0.003*"laurel" + 0.003*"stan" + 0.003*"mickey" + 0.003*"stephen" + 0.003
(1, '0.003*"cassavetes" + 0.003*"build" + 0.003*"touched" + 0.002*"damned" + 0.002*"daddy" + 0.
(2, '0.007*"plot" + 0.004*"scenes" + 0.004*"acting" + 0.004*"character" + 0.004*"watch" + 0.004
(3, '0.006*"funny" + 0.005*"original" + 0.003*"cast" + 0.003*"actors" + 0.003*"character" + 0.0
(4, '0.005*"woman" + 0.004*"cannibal" + 0.004*"david" + 0.003*"young" + 0.002*"years" + 0.002*"
(5, '0.004*"back" + 0.003*"werewolf" + 0.003*"funny" + 0.003*"disney" + 0.003*"worst" + 0.002*"
(6, '0.003*"value" + 0.003*"expecting" + 0.002*"kevin" + 0.002*"truck" + 0.002*"williams" + 0.0
(7, '0.007*"virus" + 0.005*"system" + 0.003*"drugs" + 0.003*"brainless" + 0.003*"frequently" +
(8, '0.006*"game" + 0.004*"christian" + 0.004*"janeane" + 0.003*"standard" + 0.003*"chaplin" +
(9, '0.008*"acting" + 0.007*"watch" + 0.007*"plot" + 0.006*"scene" + 0.006*"character" + 0.005*"
LDA is not required for positive reviews.
```

The Latent Dirichlet Allocation (LDA) model identified distinct topics from the negative reviews after sentiment analysis. Each topic is represented by a collection of words that frequently appear together in the reviews. These words help to form a coherent theme or subject that the reviews focus on.

For instance, some topics might revolve around specific issues such as "poor acting," "weak storyline," or "bad visual effects." These topics were automatically generated by the LDA model based on the patterns of word co-occurrence within the dataset.

3.0.3 Topics generated by LSA model

Latent Semantic Analysis (LSA) approaches topic modeling differently by capturing the underlying structure in the text data through singular value decomposition (SVD). The LSA model generates topics by identifying patterns in the relationships between terms in the dataset.

The topics generated by LSA are generally more abstract and may capture broader themes. For example, the LSA model might generate topics that encompass general sentiments like "disappointment" or "lack of engagement," which can be attributed to various specific issues such as plot, acting, or pacing.

```
Topics for Negative Reviews:
Topic 0: acting plot watch book scenes pretty life real character funny

Topic 1: series david gathering week episodes spanish hospital original tele

Topic 2: funny watch laugh find tv series extremely boring ive interesting

Topic 3: action love scenes jackie screen padrino pure masterpiece chan glac

Topic 4: book read horror blood ive bradbury reason saras witch acting

Topic 5: heat feel theme real series adventure song wonderful life audience

Topic 6: woman pretty find laugh book start dumb friends police big

Topic 7: japanese woman knotts horror makes star pretty dull roles kid
```

The below table displays the results of Topic Modeling using both LDA (Latent Dirichlet Allocation) and LSA (Latent Semantic Analysis). To evaluate the quality of the topics generated, we use a measure called the coherence score, which indicates how well the topics make sense. In this analysis, the LDA model achieved a coherence score of 0.68, while the LSA model performed better with a higher coherence score of 0.79. This suggests that LSA is more effective at identifying meaningful and coherent topics in the data compared to LDA. The coherence score is crucial because it helps determine the relevance and clarity of the topics extracted, making LSA a more reliable choice in this scenario.

Table 3.2: Results of the Topic Modelling methods

Topic Modelling	Coherence
LDA	0.68
LSA	0.7883

CHAPTER 4

CONCLUSION

The project successfully achieved its objective of extracting actionable insights from user feedback through a combination of traditional machine learning and deep learning techniques. The sentiment analysis models, particularly the Bidirectional LSTM, demonstrated high accuracy, validating the effectiveness of deep learning for text classification tasks. The topic modeling results provided further insights by uncovering specific concerns and themes prevalent in negative reviews. These insights can be used by stakeholders to address user concerns, improve services, and enhance overall user satisfaction.

The integration of LSA for topic modeling yielded a coherence score of 0.79, indicating a strong correlation between identified topics and the content of the reviews. This result underscores the importance of selecting appropriate models and techniques for different stages of text analysis.

Looking forward, the methods developed in this project can be applied to other datasets and domains, offering a scalable solution for sentiment analysis and topic modeling. Future work could explore more sophisticated models, such as Transformer-based architectures like BERT or GPT, to further enhance the accuracy and depth of analysis. Additionally, incorporating user feedback into the iterative development process could lead to more personalized and effective insights, ultimately driving better decision-making and user satisfaction.

REFERENCES

[1] Feng Li, Minlie Huang, and Xiaoyan Zhu. Sentiment analysis with global topics and local dependency. In *Proceedings of AAAI*, pages 1371–1376, 2010.

[2] Chia-Hsiu Lin, Yifan He, Richard Everson, and Simon Ruger. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, 2012.

[3] Tariq A. Rana, Y.-N. Cheah, and S. Letchmunan. Topic modeling in sentiment analysis: A systematic review. *Journal of ICT Research and Applications*, 10(1):76–93, 2016.

[4] Qiaozhu Mei, Xue Ling, Michael Wondra, Hongfang Su, and ChengXiang Zhai. Topic-sentiment mixture: Modelling facets and opinions in weblogs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 171–180, 2007.

[5] Chao Xue, Wei Tang, Hong Xu, and Xian Hu. Double lda: A sentiment analysis model based on topic model. In *Proceedings of the 2014 10th International Conference on Semantics, Knowledge and Grids*, pages 49–56. IEEE Computer Society, 2014.

[6] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Draft of August 20, 2024, 2024. Copyright © 2024. All rights reserved.